AD661973

AFOSR 67-2667

23

HERNER AND COMPANY
2431 K STREET, N. W.
WASHINGTON, D. C. 20037

TELEPHONE
965-3100

CABLE ADDRESS
"HERNCO"

October 5, 1967

Dr. Harold Wooster, Director
Information Sciences Directorate
Air Force Office of Scientific Research
(Attn. SRIR)
1400 Wilson Boulevard
Arlington, Virginia 22209

Subject: Final Scientific Report
Contract No. AF49(638)-1617

Dear Dr. Wooster:

I write to report on activities of the subject contract since its inception, October 1, 1965.

1. Notification Theory. 'Information Flow'. Formal and informal discussions of the foundations and implications of this theory have removed many ambiguities, tightened its structure, and sharpened its applicability. A paper expounding the motivation and basis of the theory will be published in J. Association for Computing Machinery, October 1967 issue.

To summarize, the scope of 'Notification' (i.e., Information Retrieval and cognate documentary activities) is delimited. All such activities are the tools of discourse. Without altering the discourse they cannot participate in it, they cannot anticipate it as authors, or as printers, signallers, or typists. Still less can they transcend discourse; e.g., by evaluating the relevance of some part of the discourse to another, or by guaranteeing that a particular record will help a particular individual. Such activities not only demand omniscience, but even then must be retrospective.

Within these limits, management of recorded discourse must cope with six variables. In broad terms these may be named Message, Channel, Code, Source, Destination, Designation. The first three are the variables of Shannon's Information Theory, the last three are the variables of Discourse; i.e., of the study of who talks to whom about what, irrespective of the language or mode of communication.

Besides these two triads, there are necessarily eighteen others. These can be identified as the atomic activities of Notification (i.e., of management of recorded discourse). Most are familiar. In theory, if the variables are strictly defined, entropic measures can be applied to all twenty, as well as to the familiar Message, Code, Channel triad of Shannon's measure of Selective Information. They do not necessarily measure 'information' in any sense. For instance, in the Source, Destination, Designation

Distribution of this document is unlimited.

triad we get a measure of complexity of discourse, in the sense of average variability of subject matter. In the Source, Destination, Channel triad, we get a measure of average unpredictability of traffic.

By themselves, these triads, and measures associated with them, do not involve any form of 'flow' explicitly. To do this, one must have two triads with two variables in common. That is, a tetrad of properly defined variables. With this configuration, a 'flow' or, at any rate, correspondence between any pair of variables entails a conjugate flow or correspondence between the other two. For instance, shelf arrangement of records by subject matter requires establishment of the triads Channel (site), Code (address), and Message, and of Designation (topic), Code (label of topic), and Message. This demands that the Codes should be, at the same time, the address of the material record on the shelves, and the code-name for its topic. This being so, the correspondence between Channel and Code (address) reflects the correspondence between Message and Designation; the correspondence between Channel (site) and Designation (topic) reflects the correspondence between Message and Code (classification number); and the correspondence between Message and Channel reflects the correspondence between Designation and Code. All these depend on establishment of two original triads so that the Codes are both the addresses and the labels for the subject matter of the same set of messages.

Unless the basic tetrad of variables are compatible in this way, no flow or correspondence is possible. When a tetrad is self-compatible, three conjugate flows are implied, but they are not independent. Because there are six variables, there are fifteen distinct types of correspondence, any of which are liable to be called 'information flow' without further explanation. Because these flows are inter-dependent, systems purporting to promote them are vulnerable to incompatibility as well as to ambiguity.

These fundamental principals can be applied to Information System design on one hand, and on the other, to examination of the logical foundations and nature of such informational activities as involve recorded discourse, which includes instrument records, photographs, and other artifacts intended to be used as records.

Wide discussion shows these principles to agree with the every day views of practical documentalists. In particular, with the 'operation' analysis of working systems, as typified by the approach of Lea Bohnert. On the other hand, it agrees with the 'behavioral' approach to problems of documentation and cognate activities.

Those who take 'knowledge' as such as the fundamental, and human use of human records as secondary phenomena, find the Notification Theory decidedly uncongenial. Why they should do so, is not obvious.

Logically, it is consistent with a Platonic view of knowledge, or with a non-Platonic view, for that matter. It simply does not concern itself with such matters, but deals with the management of recorded discourse in terms of existing records, existing production of records, and existing use of the records. It claims that all questions of 'retrieval efficiency', 'relevance', and the like, can be settled in those terms or not at all. In short, the business of 'Information Retrieval' is to provide people with what they ask for, within the limits of reasonable requests that do not involve omniscience to carry them out. At best, recorded discourse is itself a tool, and Information Systems tools for access to that tool. No tool can guarantee that it will be used sensibly, properly, or usefully. It can aim only at performing the better, the better its user.

The abilities of readers, and the improvement of their knowledge, are matters of great interest and importance. But they do not lie within the scope of systems for improving access to recorded discourse.

2. Measures of Performance and Efficiency of Retrieval Systems. Study of performance of retrieval systems is still much hampered by irrelevant considerations of 'relevance', 'user satisfaction', 'helpful to reader', and similar matters that are either unknowable or outside the competence of retrieval systems. However, even with reasonable and operational criteria for acceptability with respect to a given request, problems of measurement remain.

For some years the testing (for some reason called 'evaluation') of retrieval systems has been handicapped by two alleged measures of merit; the Cranfield Ratios, i.e., the 'relevance (precision) ratio' and 'recall ratio'. The first is the ratio of the number of retrieved and acceptable items to the total number of retrieved items; the second, the ratio of retrieved and acceptable items to the total number of acceptable items in the collection. These ratios are assumed to be fundamental characteristics of the retrieval system and have been given 'probabilistic' interpretations by statisticians who should have known better.

The behavior of a specific retrieval system with respect to specific requests is completely determinate. It may be, indeed it is, convenient to describe the results of extensive tests in terms of means, dispersions, and other statistical measures. This does not imply that retrieval is carried out, or even behaves as if carried out, by a group of little green women playing crap games. The probabilistic model implied by taking 'recall ratio' and 'precision ratio' as fundamental characteristics is fantastic. Still less does a plot of one against the other indicate anything in particular. Plots of two ratios, both of small integers, that have the same numerator are singularly uninformative.

Even if some characteristics of retrieval systems are best displayed as ratios, these certainly will not include the Cranfield Ratios. Few people would accept a response of two items, one acceptable, as of equal merit with a response of two thousand items, one thousand acceptable. Both responses have the same Precision Ratio, but in the first, one has only to pick out one acceptable item from two; in the second, one has to pick out one thousand from two thousand. Similarly, when considering the Recall Ratio, losing one item out of two is not the same as losing one thousand out of two thousand.

Equally serious, both Cranfield Ratios depend upon what is in the collection; that is, upon the habits of authors and the library acquisition policy. To see this, imagine an experiment that has yielded a response of so many acceptable and so many unacceptable items. Throw away some or all of these selected and acceptable items, and some or all of the selected and unacceptable items. Repeat the experiment on the collection so amended. Clearly, the new response will yield different ratios, though only the collection has altered. The indexing and the request formulation are unchanged.

To some extent this objection can be dodged by developing numerical measures in other terms. But the fundamental problem remains; what retrieval characteristics really reflect the merits of a retrieval method, and of the method alone? Clearly, one must regard any particular collection as being a sample of the totality of collections of documents 'of that type'. Similarly, a particular set of requests must be regarded as a sample of the totality of requests 'of that type'. To decide what are meant by 'of that type' in these two questions is fundamental.

Also, we must find out whether 'retrieval characteristics' can be separated in a meaningful way from the nature of the collection, that is, from what authors write and from library acquisition policy. Both indexing and requesting are formulated, however implicitly, in relation to the totality of items to be indexed or requested, not to each one in isolation.

One must also distinguish between the different, and sometimes incompatible, demands made on a retrieval system. In general, a reader demands at least that (1) documents of the kind requested should exist, (2) that the system should have access to them, (3) that he should be supplied with as many as possible, i.e., that the 'loss' in the response should be as small as possible, (4) that he should be supplied with as few unacceptable items as possible, i.e., that the 'padding' should be as small as possible, (5) that he be given confidence limits, or similar estimates, as to the number of acceptable items in the collection. This is usually in the form of an 'existence' request, typified by a Patent Office search. What he does not ever demand is that he be given a certain ratio, one-half or thirteen-seventeenths, say, of the acceptable items.

Clearly, (5) is of a very different kind to the other types of request; (3) and (4) cannot be satisfied simultaneously except by luck. Also, (1) and (2) are not usually regarded as 'retrieval characteristics' of a system, though retrieval is impossible unless they are fulfilled to some extent.

Thus, even if one has a rational criterion of acceptability, and has rid oneself of the more obviously erroneous numerical measures and 'models', some fundamental questions remain. In particular, in what ways, if at all, can the retrieval performance of a system be compared either with its previous performances or with other retrieval systems. It is not clear that there is such a thing as 'retrieval performance' that can be separated from other essential characteristics of record management.

It is easy enough to write a paper demolishing existing and proposed measures of retrieval performance -- including some suggestions of mine. This is useless without some solid foundation for new ones that will cope with the considerations discussed above.

I, therefore, drafted a summary of the considerations outlined above and circulated it to some of those interested in these questions. The responses varied from the rational to the emotional, according to degree of involvement with Cranfield Ratios. Fortunately, there are signs that such magic numbers are fading from fashion, and that more attention is being given to the nature of retrieval operations as revealed in practice.

3. Conferences.

3.1 Contributed papers to conferences are listed in the Appendix to this report under 'Publications'. Where the report or proceedings are still in the press, this is indicated. Contributions to discussion are listed under 'Presentations'. Usually these have been reported in full or in summary in the appropriate accounts of the meeting.

3.2 In June/July 1966, I attended three formal conferences and made several professional visits in England. The conferences and presentations are listed in the Appendix. A full account of this trip and the conclusions I draw from it, were given you in Technical Status Report No. 2, dated October 14, 1966.

In summary, I found the documentation scene in the UK depressing. Certainly the standards of criticism and understanding were much lower than in, say, 1951. Far too many people were following in each others footsteps in circles, and had been doing so for a long time. This is not unknown outside UK, but damages small countries more than it does large.

On the other hand, the UK universities shine quite brightly. This goes for old and new universities, and for traditional and non-traditional library activities.

As usual in UK, these bright spots are individuals, or unofficial associations of individuals. In general, the official outlook is benevolent, but too ignorant of the subject to tell good work from bad. The most support goes to the most noisy, and is vulnerable to fashion.

4. Committees, Consultations, Ancillary Activities.

4.1 As member of the Advisory Board, ADI Annual Reviews, I assisted with the gestation and birth of the first, 1966, volume. For this volume, I also acted as low-level referee for the contributions of Baxendale, Black, and Bourne. For the second, 1967, volume, I commented on choice of contributors, but the mechanics of the actual production of these Reviews are now almost finalized.

4.2 I continue to receive and, when appropriate, comment upon proposals of the Terminology Committee of the British Computer Society. This committee reviews, constructs, and recommends amendments to the *IFIP Vocabulary of Information Processing*. It works in conjunction with the British Computer Society as a whole, the British Standards Institution, and the International Federation for Information Processing (IFIP).

Attempts to create a rational outlook on terminology, let alone to create a rational terminology in the 'information' field or fields, have at the moment much in common with efforts to clean up a monkey-house with a single piece of Kleenex. Nevertheless, although present efforts may seem hopeless, they will provide a clean foundation for the future.

4.3 I was a minor member of the Special Activities section of the Organizing Committee for the 20th Anniversary Conference of the Association for Computing Machinery. In this capacity, I had to trace the existence and whereabouts of members of the First Executive Council (1947) and Past Presidents of the Association, and then lure them as guests to the Conference. This research resulted in a pleasantly high yield.

4.4 *Science* (organ of the AAAS) has sent me for comment several papers submitted for publication, on topics concerned with informational activities. *American Documentation* sends me papers for comment and refereeing. *Computing Reviews* sends me published papers for signed review. Individual authors sometimes send manuscripts to me directly for comment. These I deal with as time, and competence, permit.

4.5 Throughout the period of this Contract, frequent formal and informal consultations have taken place on the AFOSR/CEIR Mon Doc project. These have proved essential, because both this Contract and the project deal with unification of documentary processes. Thus, if they are properly based, they should agree closely in principle, though differing in emphasis and exposition. They do so agree. Wherever in discussions with Mrs. Bohnert, Calvin Mooers, and John O'Connor, disagreement appears, this is due to differences about the scope and nature of documentary procedures, rather than about the operational issues involved.

4.6 There have been informal contacts and correspondence with Carlos Cuadra, J. O'Connor, Alan Rees, Don Hillman, Gerard Salton, Cyril Cleverdon, and others on the interdependent issues of 'relevance', retrieval tests, retrieval performance, and numerical measures thereof. This included some criticisms of the Cranfield measures which were and are not received kindly by their proponents.

4.7 The Russian members of FID/CR circulated some papers, in English translation, advocating the label "Informatics" for the combined fields of symbol manipulation, recorded discourse, and communication methods. Or so I understand the proposal. Provided the scope is defined, the actual label used is immaterial, so long as it is internationally unambiguous. However, to me the Russians had not made the scope clear, amongst other things confusing physical entropy, entropic measures, and recorded discourse. I summarized my views in a letter sent to B. Adkinson of NSF, and President of FID, at his request.

4.8 I have paid several visits to Dr. Altman of the STINFO Library at the Harry Diamond Laboratories, and have discussed various practical and theoretical matters with him and his associates.

4.9 The Encyclopedia Britannica requested me to write a Historical Sketch for the main section 'Information Processing' that is to appear in their 1968 edition. The sketch was to cover the social, rather than technical, development of symbol manipulating (i.e., computing) devices through the pioneering electronic computers, in some 1500 words. Whether this task is possible or not, I made the attempt, hoping that references to other articles might fill the gaps. One by-product of the endeavor was the discovery that Taylor, the inventor of the 'Peek-a-Boo' system for retrieval by joint attributes, had the given name of 'Horace'.

5. Presentations, Publications. The Appendix to this report lists these.

Very sincerely yours,

HERNER AND COMPANY

R.A. Fairthorne

R. A. Fairthorne VCK

Enclosure - Appendix 1

APPENDIX 1

Contract AF49(638)-1617

PUBLICATIONS AND PRESENTATIONS
October 1965 through September 1967

1. Publications

(abstract) Notification Theory. International Federation of Documentation, (FID), and ADI, Conference Abstracts. p. 66, Oct. 1965

Some Basic Comments on Retrieval Testing. J. Documentation 21, 4, pp. 267-270, December 1965

(letter) Who Pilots the Hovercraft? J. Documentation, 22, 2, p. 46, June 1966

Morphology of 'Information Flow'. J. Association for Computing Machinery, October 1967

'The Applied Mathematics of H. P. Luhn', in: Schultz, C.K. (ed), Hans Peter Luhn - Pioneer and Prophet of Information Processing. (in the press)

Information Processing: Historical Sketch. Encyclopedia Britannica, 1968 (in the press)

Critique of Borko's 'Conceptual Foundations', in Foundations of Access to Knowledge; a Symposium, July 1965, Syracuse University. (to be published)

Critique of Soergels' 'Remarks on Information Languages', in International Symposium on Relational Factors in Classification, June 1966, University of Maryland (to be published)

2. Presentations

Presentation of Paper at and participation in

FID/ADI Congress, Washington, D.C. 10-15 October 1965

American University, Center for Technology and Administration, Course on Management of Technical Records, Address on 'Subject Headings v. Descriptors' 20 Jan.1966

University of Maryland, School of Library and Information Servi[illegible]s. Colloquium on "Notification Theory". 23 Mar. 1966

Critique of papers, chairman at some sessions and participant in

International Symposium on Relational Factors in Classification, University of Maryland. 8-11 June 1966

Research Analysis Corporation, Library and Logistics Staff. Informal discussion on fundamentals of Information Retrieval. 15 June 1966

Participation, by invitation, in

ASLIB Conference on Computer Applications in Public Libraries, London, England. 21 June 1966

City University, London, England, Information Sciences graduate class. Colloquium on fundamentals of Information Retrieval. 23 June 1966

Participation, by invitation, in

Anglo-American Conference on Mechanization of Library Services, Brasenose College, Oxford, England. 30 June-3 July 1966

National Physical Laboratory, Teddington, England. Autonetics Division. Address on Notification Theory and 'Information Flow'. 6 July 1966

Inst. of Information Scientists Conference, Jesus College, Oxford, England. 11-13 July 1966

University of Maryland, School of Library and Information Services. Address on Structure of Information Activities. 21 March 1967

State University of New York at Albany, School of Library Science. Address on The Roles of Theory and Practice in Information Work. 20 May 1967

Organizing Committee, Special Events. Assoc. Computing Machinery, 20th Anniversary Meeting, Washington, D.C. 29-31 August 1967

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)
Herner and Company, 2431 K Street, N.W. (Suite 100)
Washington, D.C. 20037

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
UNIFICATION OF THEORY AND EMPIRICISM IN INFORMATION RETRIEVAL

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Final

5. AUTHOR(S) (First name, middle initial, last name)
Robert A. Fairthorne

6. REPORT DATE
5 October 1967

7a. TOTAL NO. OF PAGES
9

7b. NO. OF REFS
8

8a. CONTRACT OR GRANT NO.
AF 49(638)-1617

b. PROJECT NO.

c. 61445014

d. 681304

9a. ORIGINATOR'S REPORT NUMBER(S)

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
AFOSR 67-2667

10. DISTRIBUTION STATEMENT
1. Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Air Force Office of Scientific Research
Directorate of Information Sciences
Arlington, Virginia 22209

13. ABSTRACT

This report outlines and comments upon: 1) Notification Theory and 'Information Flow'. Notification activities are those that notify or deliver recorded messages to readers at the reader's request. It is shown that they involve twenty elementary activities. These correspond to the triads of the six variables - Source, Destination, Designation, Message, Channel, Code. 'Flow' between two of these variables exists only when they belong to a tetrad formed from two triads with another two variables in common. Thus, six distinct flows can exist.
2) Measures of Retrieval Efficiency. 'Precision Ratio' and 'Recall Ratio' are meaningless, because they can be altered at will by altering the document collection alone. Also, they imply a bizarre 'probabilistic' model of retrieval, and of readers' requirements. Before even the existence of numerical measures of retrieval efficiency can be accepted, certain problems must be solved.
3) Conferences, Committees. The author comments on conferences and committees, USA and UK, which he attended 1966-67.

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information Information Retrieval Information Sciences Information Systems Notification Recall Ratio Relevance Ratio Retrieval Efficiency Terminology						